A Service-Oriented Knowledge Management Framework over Heterogeneous Sources

Larry Kerschberg
E-Center for E-Business
George Mason University
http://eceb.gmu.edu/

NASA IS&T Colloquium Series - March 10, 2004

Outline of Presentation

- Organizational Drivers for Knowledge Management
- Technological Drivers
- Ontologies and Knowledge Organization
- Intelligent Web Search - *WebSifter*
- Agent-Based Search over Heterogeneous Sources - *Knowledge Sifter*
- Service-Oriented Knowledge Management Framework
- Conclusions, Future Work and Questions

KM Organizational Drivers

- The management of organizational ***knowledge resources*** is crucial to maintaining ***competitive advantage,***
- Organizations need to motivate and enable their ***knowledge workers*** to be more productive through ***knowledge sharing and reuse,***
- Organizations are ***outsourcing knowledge creation*** to external companies, so knowledge stewardship is important,
- Knowledge is also being created globally, so that we need to ***search for knowledge*** relevant to the enterprise.
- The *Internet* and the *Web* are ***revolutionizing*** the way an enterprise does business, science and engineering!
- Intellectual Property over the Internet Protocol (IP over IP)

Confluence of Technology Drivers

- **Web Services**
  - Enabling computer-to-computer information processing via enhanced protocols based on HTTP
  - Standards such as XML, SOAP, WSDL and UDDI
- **Semantic Web & Semantic Web Services**
  - Bringing meaning, trust and transactions to the Web
  - Creating an object-oriented Web information space
  - Standards such as Web Ontology Language (OWL)
- **GRID Services**
  - Regarding computing as an information utility
  - Custom configure remote computing dynamically
- **Service-Oriented Architectures**
  - Providing computing and information processing as services
  - Software agents to manage services

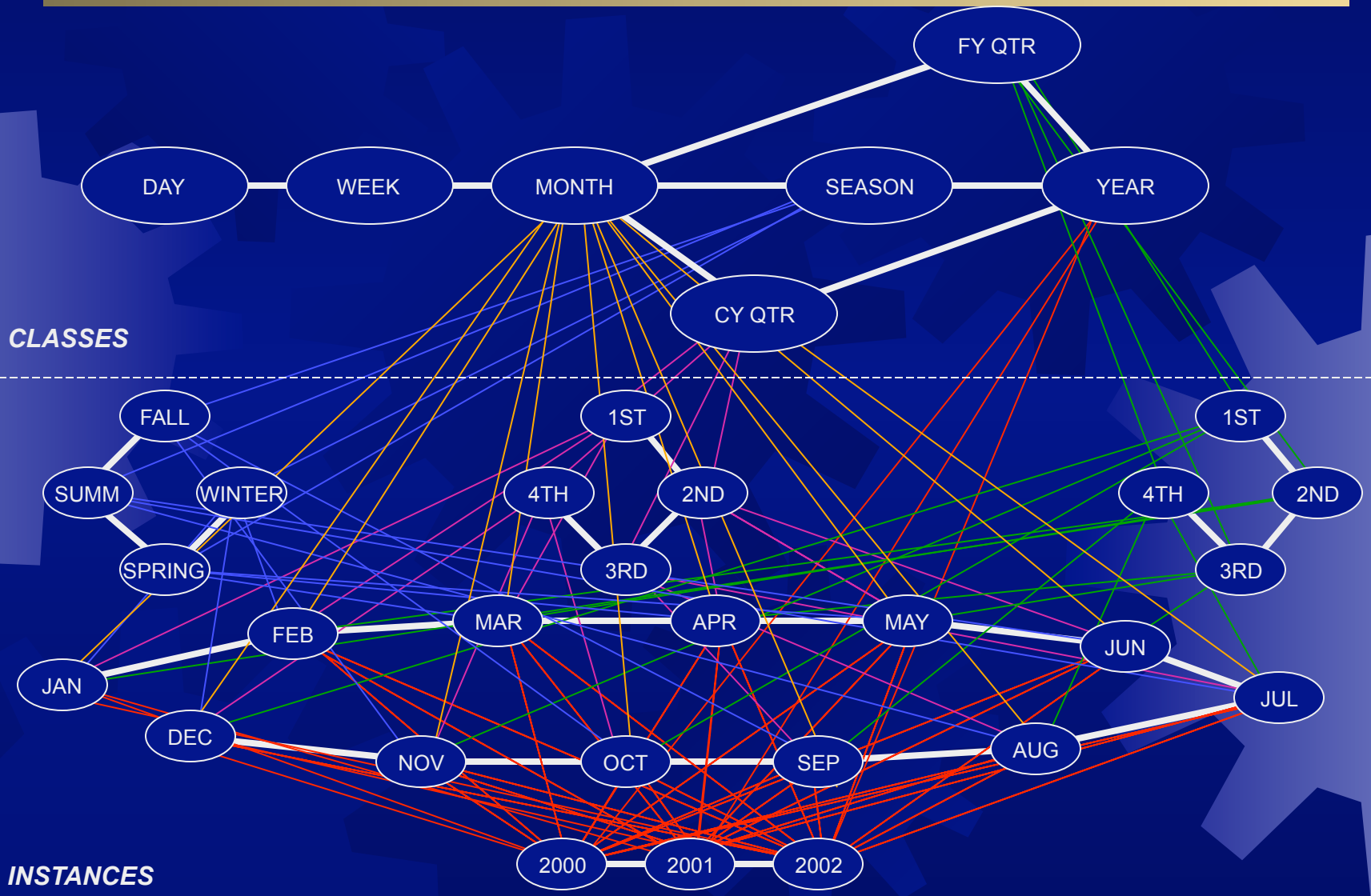
Ontology and Knowledge Organization

- **"An ontology is a formal explicit specification of a shared conceptualization" (Tom Gruber, 1993)**
  - *Conceptualization* is an abstract simplified view of the world
  - *Specification* represents the conceptualization in concrete form
  - *Explicit* because all concepts and constraints used are explicitly defined
  - *Formal* means an ontology should be machine understandable
  - *Shared* indicates the ontology captures consensual knowledge

Principles of Ontology (John Sowa)

- An *ontology* is a catalog of the types of things that are assumed to exist in a domain of interest
- Types in the ontology represents predicates, word senses, or concept and relation types
- Un-interpreted logic, such as predicate calculus, conceptual graphs, or Knowledge Interchange Format (KIF), is *ontologically neutral*.
- Logic + Ontology = language that can express relationships about entities in the domain of interest

Temporal Ontology



Taxonomic Knowledge Organization

- Service-Oriented Knowledge Management
- Taxonomic Category Pathways
  - Service-oriented Knowledge Management
  - Semantic Web
    - http://directory.google.com/Top/Reference/Knowledge_Management/Knowledge_Representation/Semantic_Web/?il=1
  - Semantic Web Taxonomy:
    - Reference > Knowledge Management > Knowledge Representation > Semantic Web
      Related Category:
      Reference > Libraries > Library and Information Science > Technical Services > Cataloguing > Metadata
  - Go to Directory Home
  - Published Ontologies on Goggle
  - JPL Semantic Web for Earth and Environmental Terminology

WebSifter II: A Semantic Taxonomy-Based Personalizable Meta-Search Agent

Larry Kerschberg, E-Center for E-Business, George Mason University (http://eceb.gmu.edu/)

Wooju Kim, Chonbuk National University, Korea, GMU Visiting Scholar.

Anthony Scime, SUNY- Brockport

Limitations of Search Engines

- **Web Coverage of Search Engines**
  - By Steve Lawrence and C. Lee Giles (July 1999)
  - The best existing search engine covered only 38.3% of the indexable pages.
  - This motivates the need for Meta-Search Engines.
- **Weakness in Query Representation**
  - Limited to keyword-based query approach.
  - This query representation is insufficient to express fully a user's intent, as motivated by a complex problem.

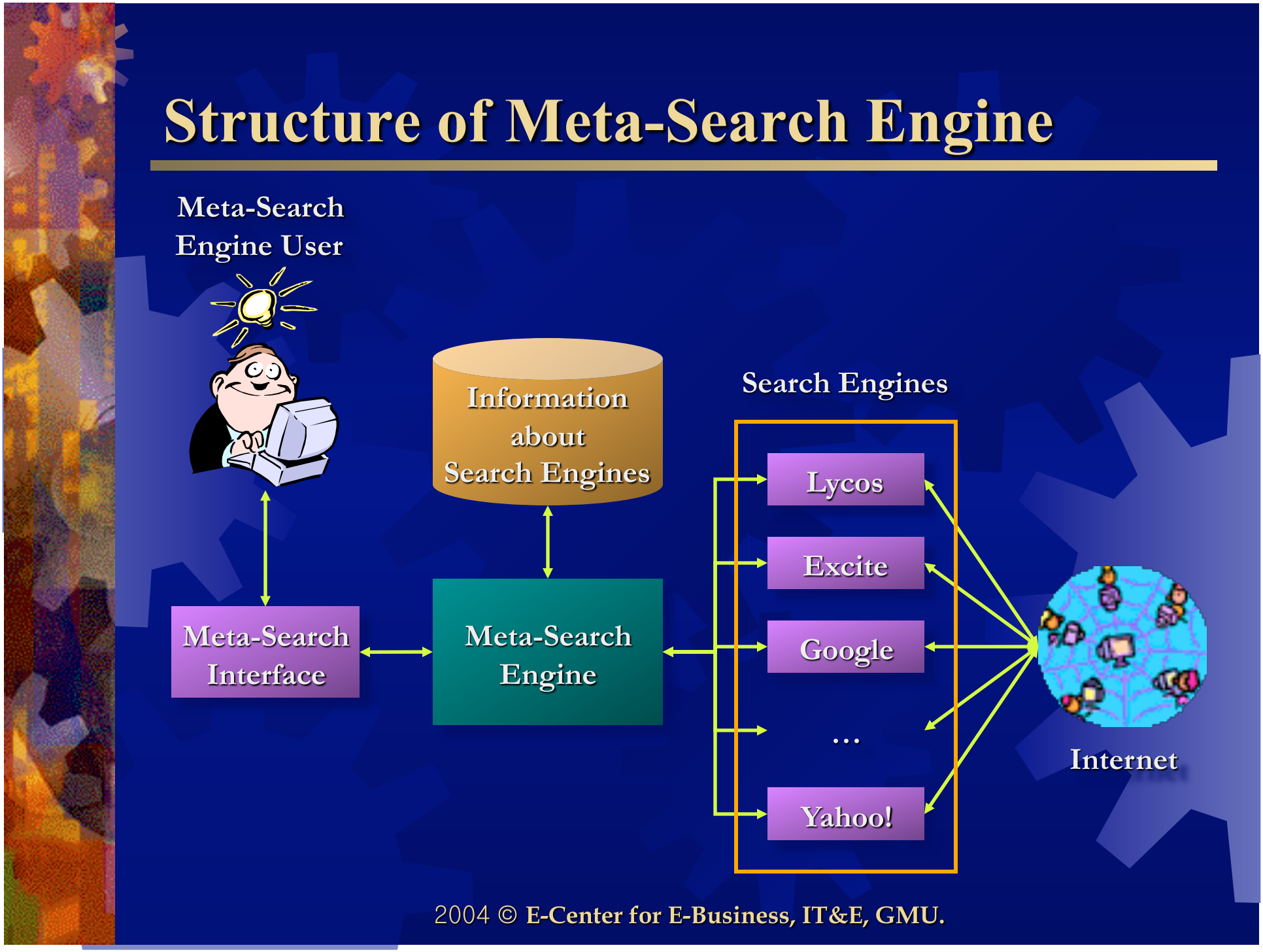
Limitations of Search Engines (Cont'd)

- **Semantic Gap**
  - Words usually have multiple meanings.
  - Most current search engines cannot identify the correct meaning of a word, and certainly not the users' intent.
- **Example by S. Chakrabarti et al. (1998)**
  - `jaguar speed' query by a wildlife researcher results in:
    - Car, Atari video game, Apple OS X, LAN server, …
    - Google Search for Jaguar Speed
    - Google Search for Animal Jaguar Speed

Limitations of Search Engines (Cont'd)

- **Lack of Customization in Ranking Criteria**
  - Users cannot personalize a search engine with their preferences regarding search criteria and/or search attributes
    - Most search engines have their own proprietary search criteria and ranking criteria.
    - For a shopping agent, lowest price may be one of many decision variables, including stock availability, flexible return policy and delivery options, return policy, etc.
  - We would like to enrich search evaluation criteria to capture user preferences regarding page ranking, including:
    - semantic relevance,
    - syntactic relevance - page location in the web structure,
    - category match,
    - popularity, and
    - authority/hub ranking.

Structure of Meta-Search Engine

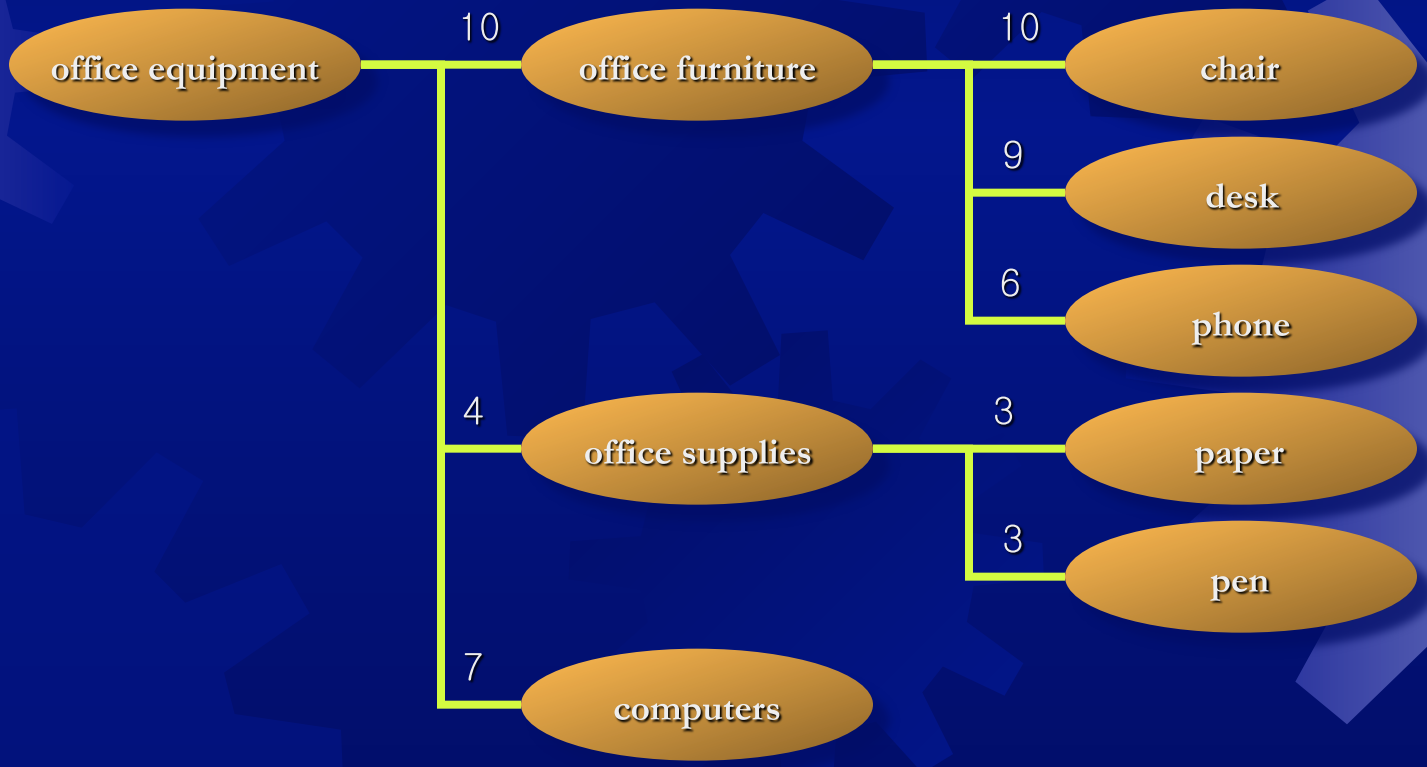


Semantic Taxonomy-Tree Approach for Personalized Information Retrieval

- **WebSifter overcomes the limitations of current search engines:**
  - Weak representation of user's search intent
  - Semantic gap of word meanings, and
  - Lack of user-specified search ranking options
- **WebSifter approach consists of:**
  - Weighted Semantic Taxonomy Tree query representation
  - Positive and negative concept identification using an ontology service
  - Search preference component selection and weighted component ranking scheme

Weighted Semantic Taxonomy Tree (WSTT)

- **Full example of a businessman's problem**
  - In WSTT, user can assign numerical weights to each concept, thereby reflecting user-perceived relevance of the concept to the search.



Semantic Considerations in WSTT

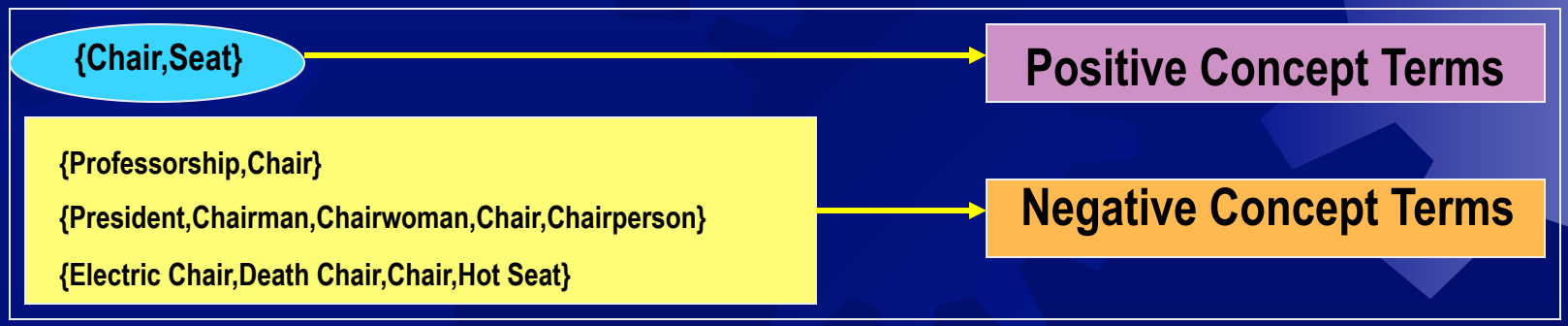
- **Multiple Meanings of a Term**
  - A term in English usually has multiple meanings and this is one of the major reasons that search engines return irrelevant search results.
- **WordNet (G. A. Miller, 1995)**
  - WordNet® is an on-line linguistic database (an on-line ontology server) where English nouns, verbs, adjectives and adverbs are organized into synonym sets (synsets), each representing one underlying lexical concept.
  - We rename this synset as Concept.
  - Thus, WordNet® provides available concepts for a term, thereby allowing users to focus on the proper search terms.

Concept Selection in WSTT

- **Example Concepts for "chair" from WordNet**
  - {chair, seat}
    - A seat for one person, with a support for the back
  - {professorship, chair}
    - The position of professor, or a chaired professorship
  - {president, chairman, chairwoman, chair, chairperson}
    - The officer who presides at the meetings of an organization
  - {electric chair, chair, death chair, hot seat}
    - An instrument of execution by electrocution; resembles a chair
- **Concept Selection for "chair"**
  - Select one among those available concepts for "chair".
  - We consider the remaining concepts as a negative indicator of user's search intent.

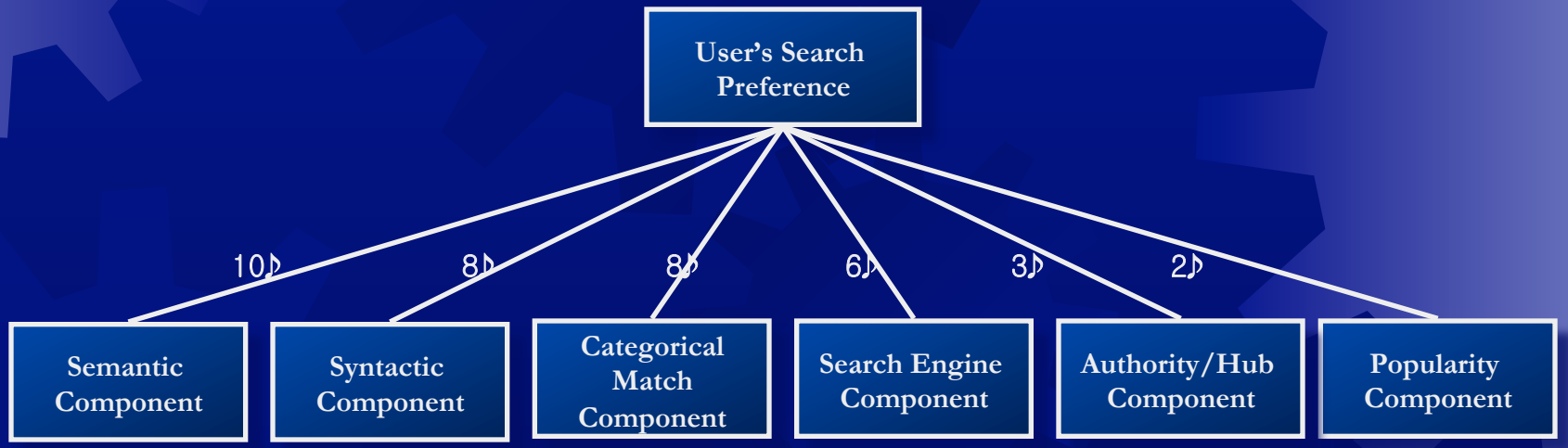
Transformed Queries for Traditional Search Engines

- **Example of Translation Mechanism**
  - For a path of WSTT such as {office → furniture→ chair}
  - Generated Boolean queries from the nodes in the path:
    - "office" AND "furniture" AND "chair"
    - "office" AND "furniture" AND "seat"
    - "office" AND "piece of furniture" AND "chair"
    - "office" AND "piece of furniture" AND "seat"
    - "office" AND "article of furniture" AND "chair"
    - "office" AND "article of furniture" AND "seat"



Search Preference Representation (1)

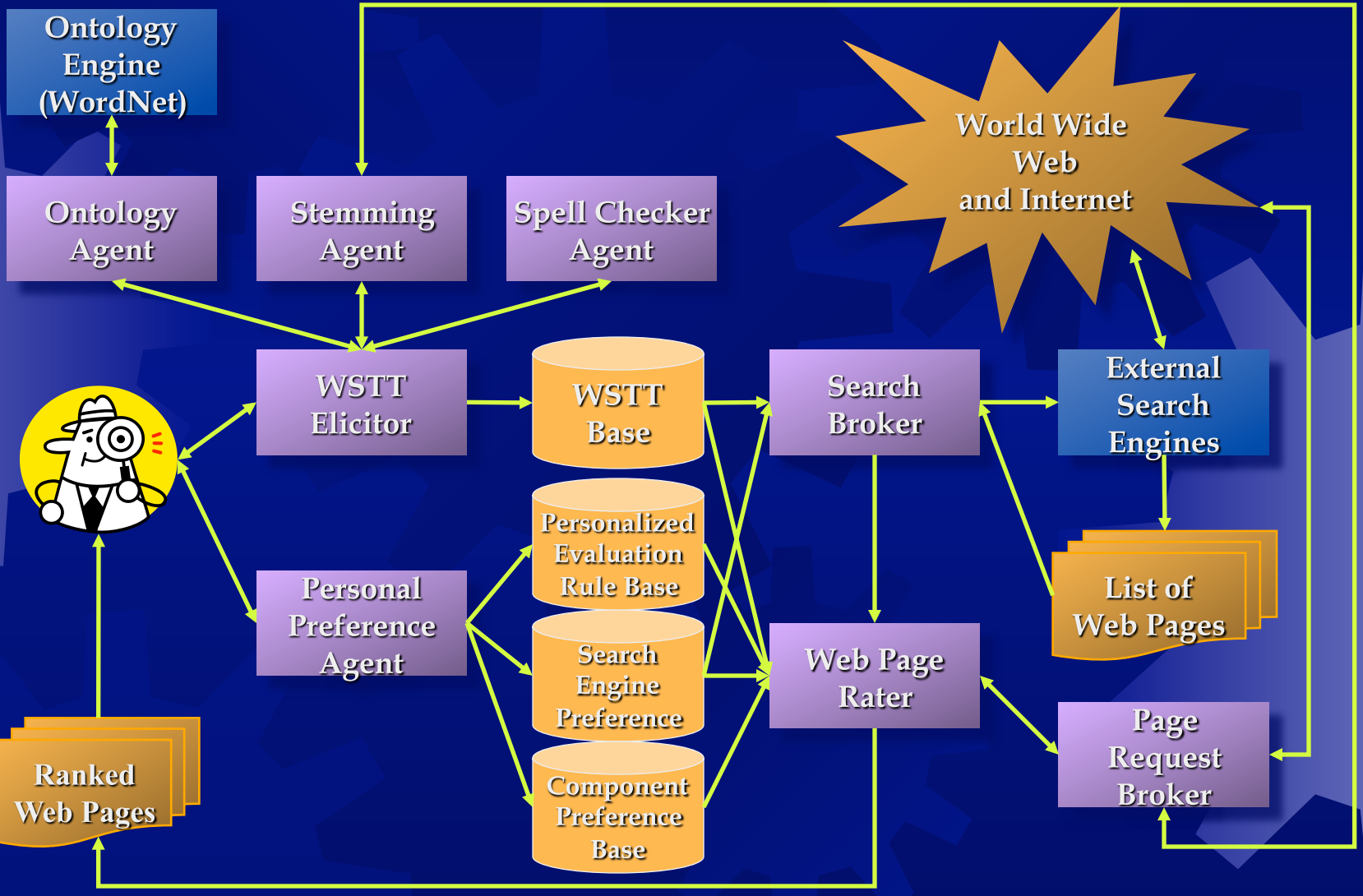
- **Preference Representation Scheme**
  - WebSifter provides a search preference representation scheme that combine both decision analytic methods,
    - MAUT (D. A. Klein, 1994) and
    - Repertory Grid (J. H. Boose and J. M. Bradshaw, 1987).
  - Component-based Preference Representation



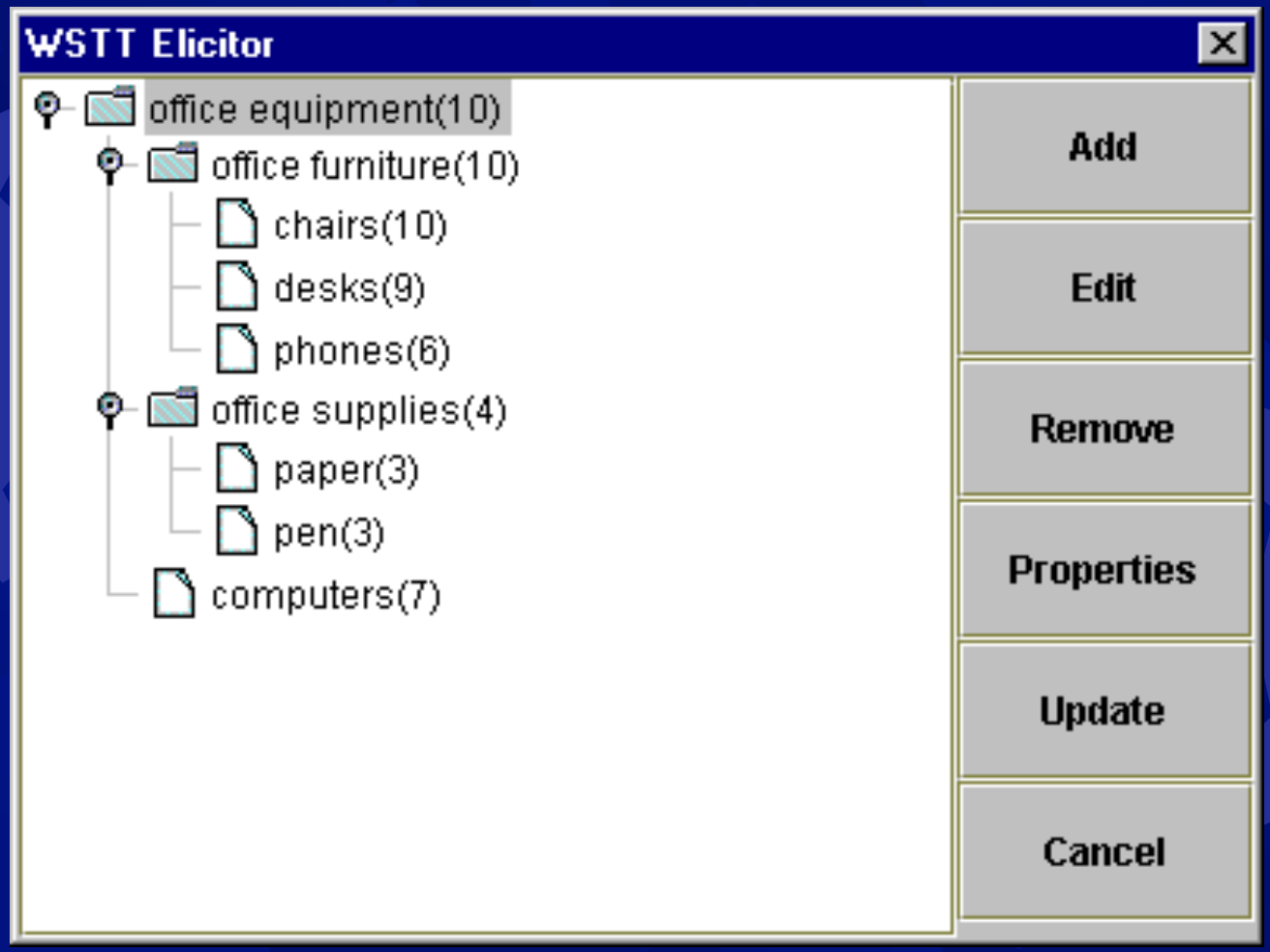
Search Preference Representation (2)

- **Six Search Preference Components**
  - *Semantic component*: represents a Web page's relevance with respect to its content.
  - *Syntactic component*: represents the syntactic relevance with respect to its URL. This considers URL structure, the location of the document, the type of information provider, and the page type (e.g., home, directory, and content).
  - *Categorical Match component*: represents the similarity measure between the structure of user-created WSTT taxonomy and the category information provided by search engines for the retrieved Web pages.
  - *Search Engine component*: represents the user's biases toward and confidence in a search engine's results.
  - *Authority/Hub component*: represents the level of user preference for Authority or Hub sites and pages.
  - *Popularity component*: represents the user's preference for popular sites.

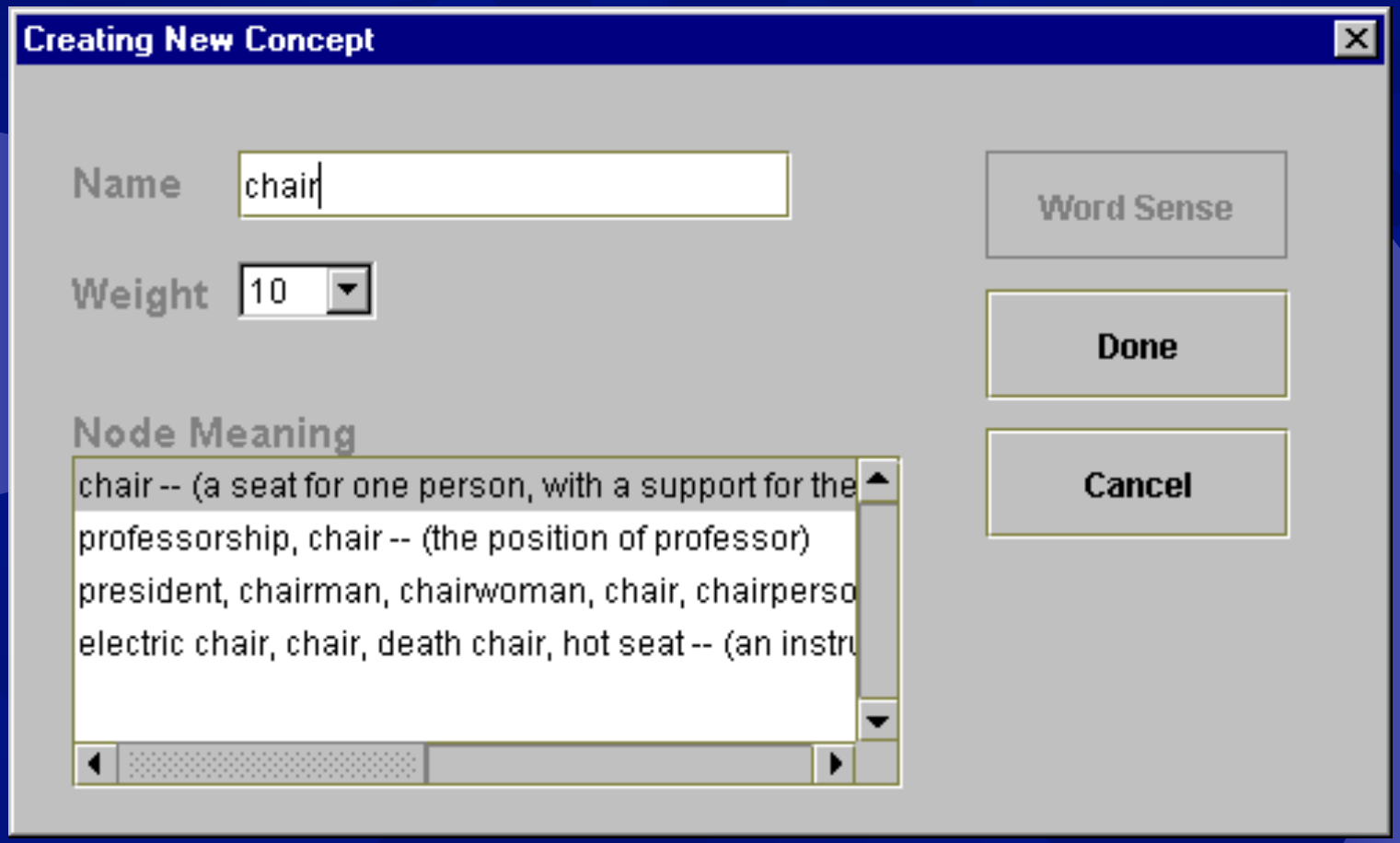
WebSifter Conceptual Architecture



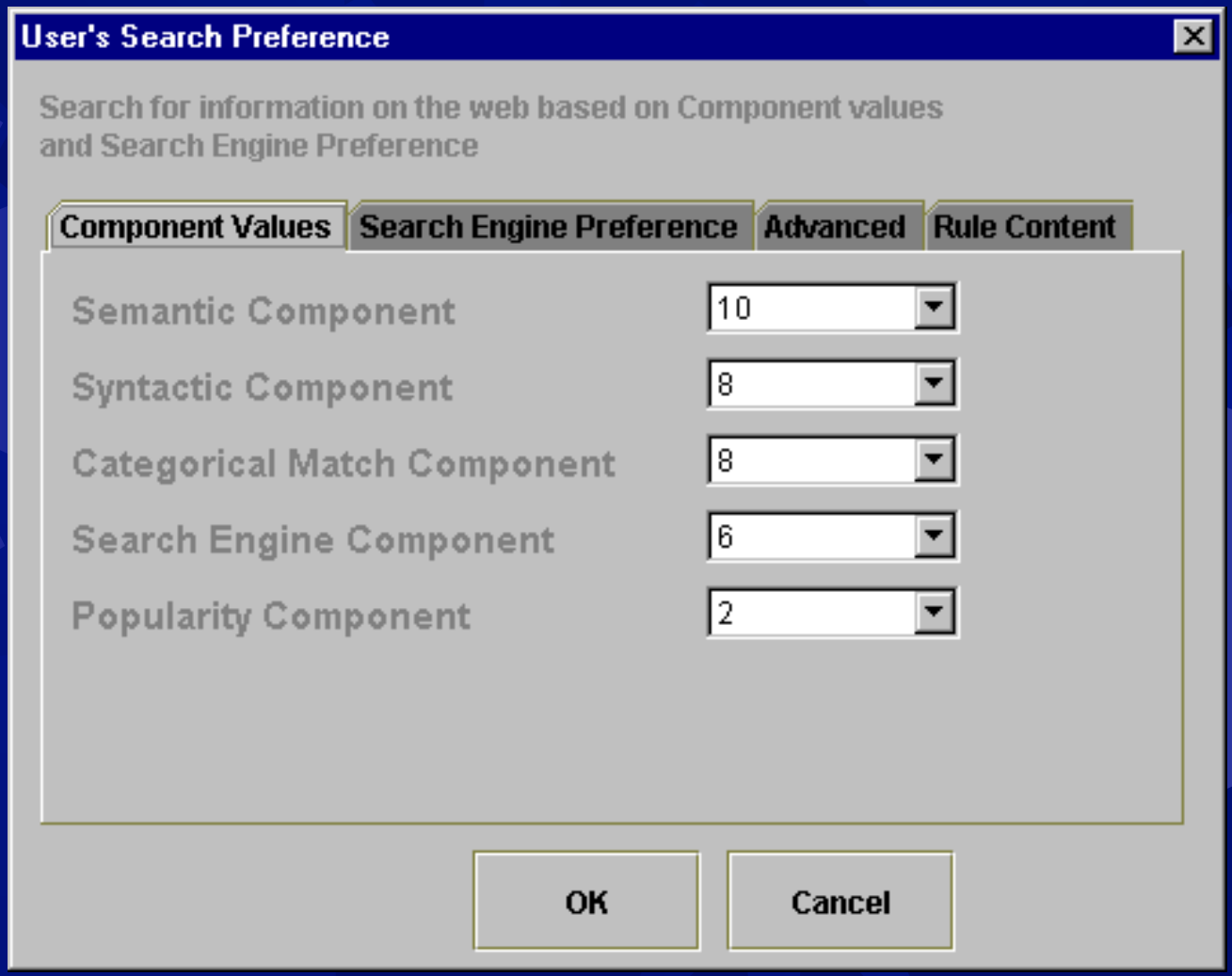
System Screen Shots – WSTT Elicitor



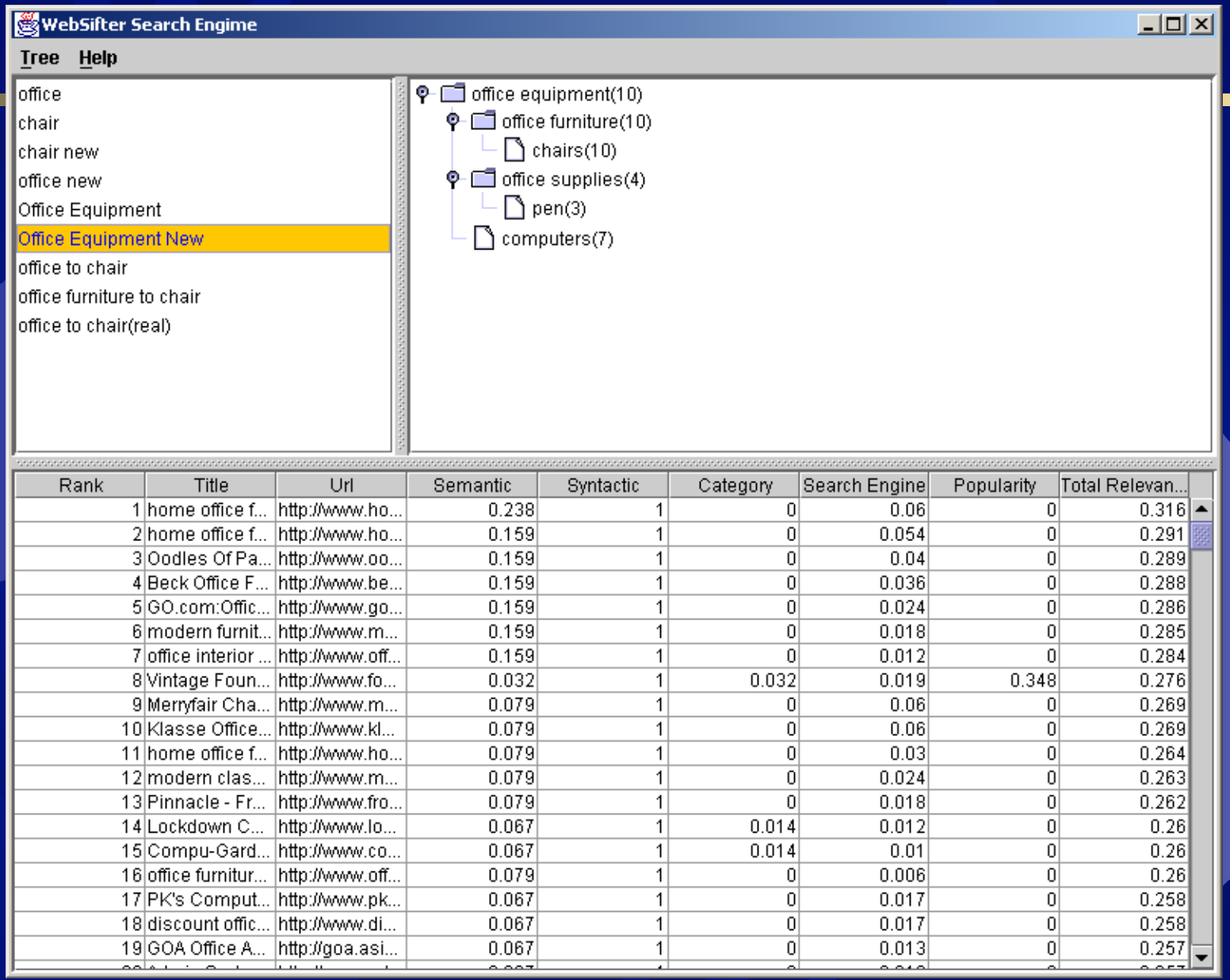
Screen Shots – Concept Selection



Screen Shot – User Search Preferences



WebSifter Main Screen



Rank	Title	Url	Semantic	Syntactic	Category	Search Engine	Popularity	Total Relevan...
1	home office f...	http://www.ho...	0.238	1	0	0.06	0	0.316
2	home office f...	http://www.ho...	0.159	1	0	0.054	0	0.291
3	Oodles Of Pa...	http://www.oo...	0.159	1	0	0.04	0	0.289
4	Beck Office F...	http://www.be...	0.159	1	0	0.036	0	0.288
5	GO.com:Offic...	http://www.go...	0.159	1	0	0.024	0	0.286
6	modern furnit...	http://www.m...	0.159	1	0	0.018	0	0.285
7	office interior ...	http://www.off...	0.159	1	0	0.012	0	0.284
8	Vintage Foun...	http://www.fo...	0.032	1	0.032	0.019	0.348	0.276
9	Merryfair Cha...	http://www.m...	0.079	1	0	0.06	0	0.269
10	Klasse Office...	http://www.kl...	0.079	1	0	0.06	0	0.269
11	home office f...	http://www.ho...	0.079	1	0	0.03	0	0.264
12	modern clas...	http://www.m...	0.079	1	0	0.024	0	0.263
13	Pinnacle - Fr...	http://www.fro...	0.079	1	0	0.018	0	0.262
14	Lockdown C...	http://www.lo...	0.067	1	0.014	0.012	0	0.26
15	Compu-Gard...	http://www.co...	0.067	1	0.014	0.01	0	0.26
16	office furnitur...	http://www.off...	0.079	1	0	0.006	0	0.26
17	PK's Comput...	http://www.pk...	0.067	1	0	0.017	0	0.258
18	discount offic...	http://www.di...	0.067	1	0	0.017	0	0.258
19	GOA Office A...	http://goa.asi...	0.067	1	0	0.013	0	0.257

WebSifter Conclusions

- **WebSifter is an agent-based meta-search engine that enhances a user's search request via pre- and post-search processing:**
  - *Problem-solving intent* captured via Weighted Semantic Taxonomy Tree,
  - Agent-based brokered consultation with the Web-based ontology service, WordNet, to *enhance the semantics* of search request,
  - Consultation with leading Search Engines such as Google, Yahoo!, Excite, Altavista, and Copernic,
  - Web page ranking based on *user-specified relevance* components including semantic, syntactic, category, authority, and popularity.

Knowledge Sifter: Ontology-Based Search over Corporate and Open Sources using Agent-Based Knowledge Services

Dr. Larry Kerschberg

Dr. Daniel Menascé

E-Center for E-Business
http://eceb.gmu.edu/

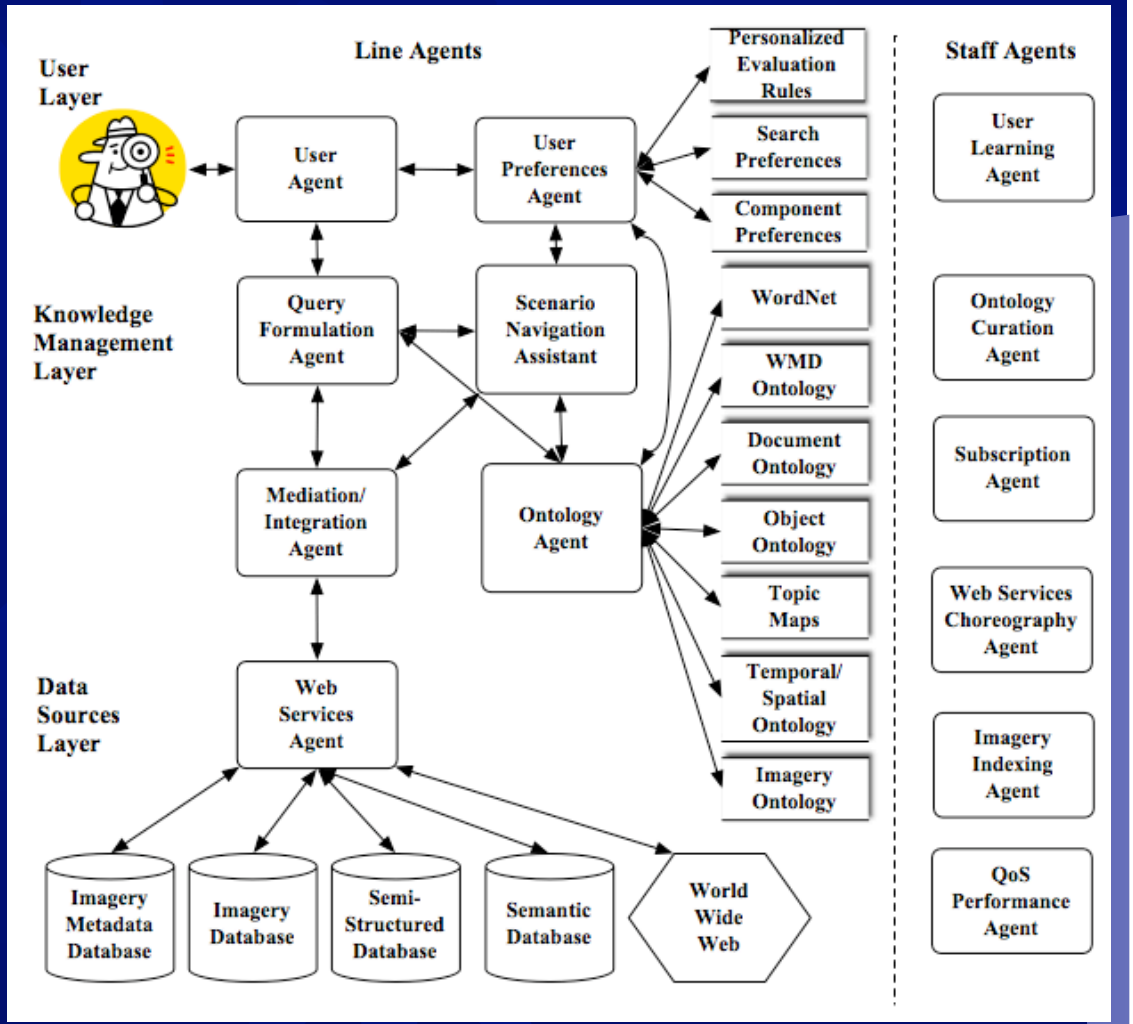
Sponsored NURI by National Geospatial-Intelligence Agency (NGA)

Knowledge Sifter Goals

- **Investigate, design and build Knowledge Sifter:**
  - An agent-based multi-layered system;
  - Based on open standards;
  - Supports analyst search, knowledge capture, and knowledge evolution.
- **Support intelligence analysts in searching for knowledge from multiple heterogeneous information sources,**
  - Use multiple, "lightweight" domain ontologies to assist analysts in posing "semantic" queries
- **Process semantic queries by decomposing them into subqueries for searching and retrieving information from multiple sources:**
  - World Wide Web, Semantic Web, XML-databases, Image Databases, and Image Metadata;

Knowledge Sifter Architecture

- KS has both line and staff agents that cooperate in managing workflow.
- User agent interacts with user to obtain preferences and search intent.
- Query formulation agent consults ontology agent to create a semantic query.
- Mediation/Integration agent decompose query into subqueries for target sources.
- Web services agent coordinates processing of subqueries.
- Staff agents work in background providing knowledge services such as QoS Performance, Indexing and Ontology Curation.



Knowledge Sifter: User Layer

- **User Agent**
  - Interacts with analyst to obtain information;
  - Cooperates with User Preference Agent to provide personalized criteria for search preferences, authoritative sites, and result ranking evaluation rules;
  - Cooperates with Query Formulation Agent to convey user preferences and the "problem" to be solved.
  - User Learning Agent (staff agent) works in the background to learn and evolve user preferences, based on feedback mechanisms.

KS: Knowledge Management Layer

- Query Formulation Agent consults the Ontology Agent to assist in specifying "semantic" queries.
- Ontology Agent interacts with multiple ontologies to specify semantic search concepts.
- Mediation/Integration Agent:
  - Receives the semantic query;
  - Decomposes it into subqueries targeted for the heterogeneous sources;
  - Submits the subqueries to Web Services Agent for processing
  - Results returned from Web Services Agent are integrated and delivered for presentation to the Analyst.
- Staff agents play important roles in Web Services Choreography, QoS Performance, User Learning, Ontology Curation, Standing Subscriptions, and Indexing.

Knowledge Sifter: Data Layer

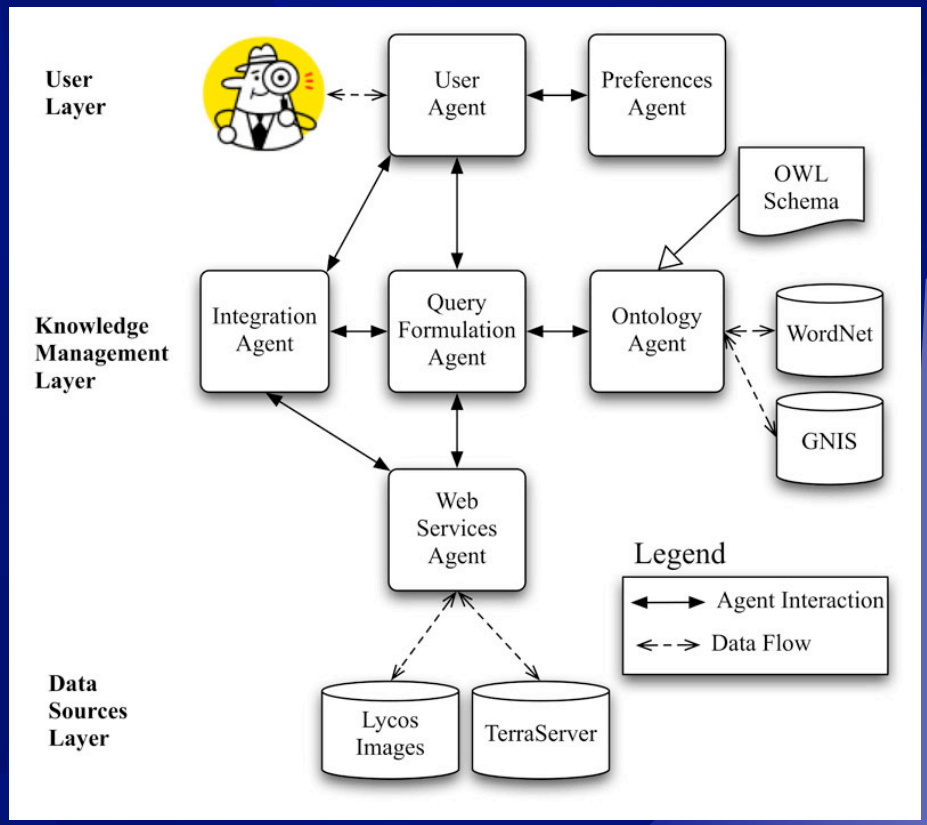
- Use of Web Services to link data source agents
- Support for heterogeneous data sources including,
  - image metadata, image archives,
  - XML-repositories,
  - relational databases,
  - the Web and
  - the emerging Semantic Web.
- Sources can register with Knowledge Sifter and begin sharing data and knowledge.
- Quality of Service Issues
  - Specification of performance and availability QoS goals.
  - QoS negotiation protocols.
  - Hierarchical caching to support scalability.

Web Services Choreography and QoS Performance Agents

- **Web Services Choreography Agent**
  - Determines composition of Web Services needed to satisfy the query
  - Builds candidate query processing plans.
  - Evaluates and decides on a plan based on user requirements
  - Implementation of response time variance reduction techniques through predictive pre-fetching, data replication, and data abstraction
- **Quality of Service Performance Agent**
  - Scalable QoS (response time and availability) monitoring of Data Layer Web Services.
  - Monitoring activity has to be adaptive to intensity of data source usage
  - Model-based performance prediction in support of Web Services Choreography agent.

Knowledge Sifter Proof-of-Concept

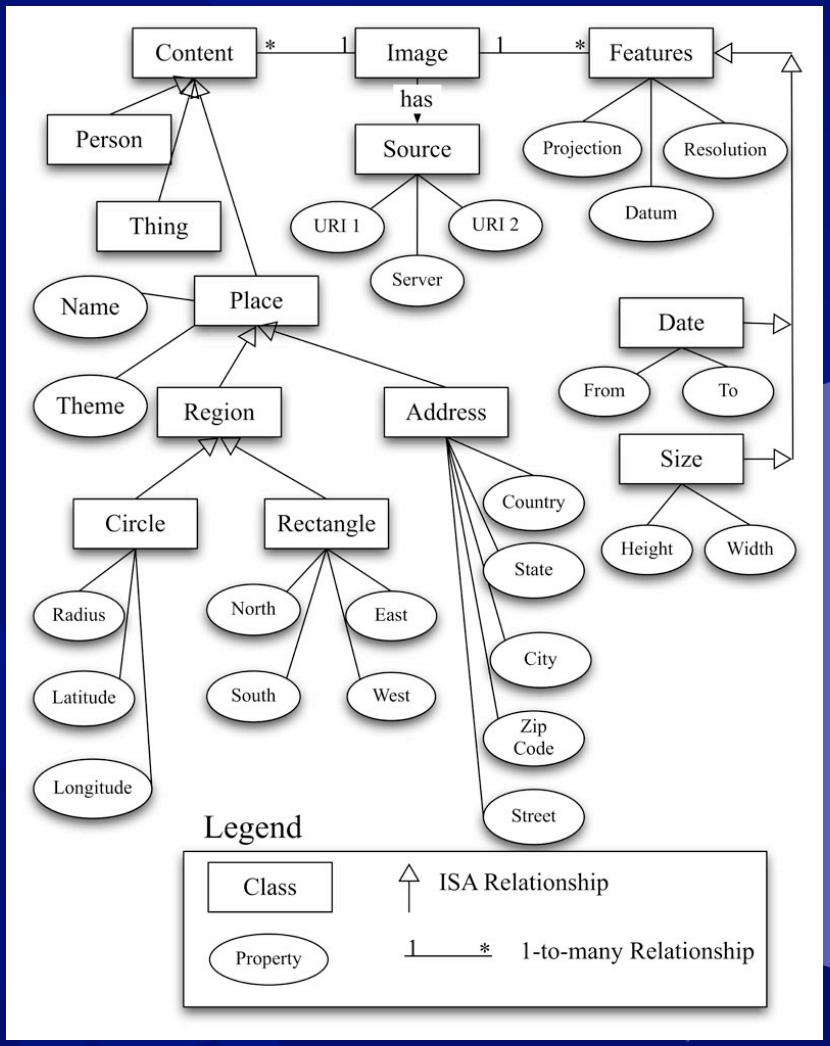
- Three-layer agent-based Semantic Web services architecture
- Ontology agent consults both WordNet and USGS's Geographic Names Information System (GNIS)
- Ontology agent conceptual model specified in Web Ontology Language (OWL)
- OWL schema instantiated by a user query, and XML-based metadata and data travel from agent to agent for lineage annotations.
- Lycos Images and TerraServer are the heterogeneous data sources.
- All agents are Web services.



Kerschberg, L., Chowdhury, M., Damiano, A., Jeong, H., Mitchell, S., Si, J. and Smith, S., Knowledge Sifter: Ontology-Driven Search over Heterogeneous Databases. (Submitted for Publication)

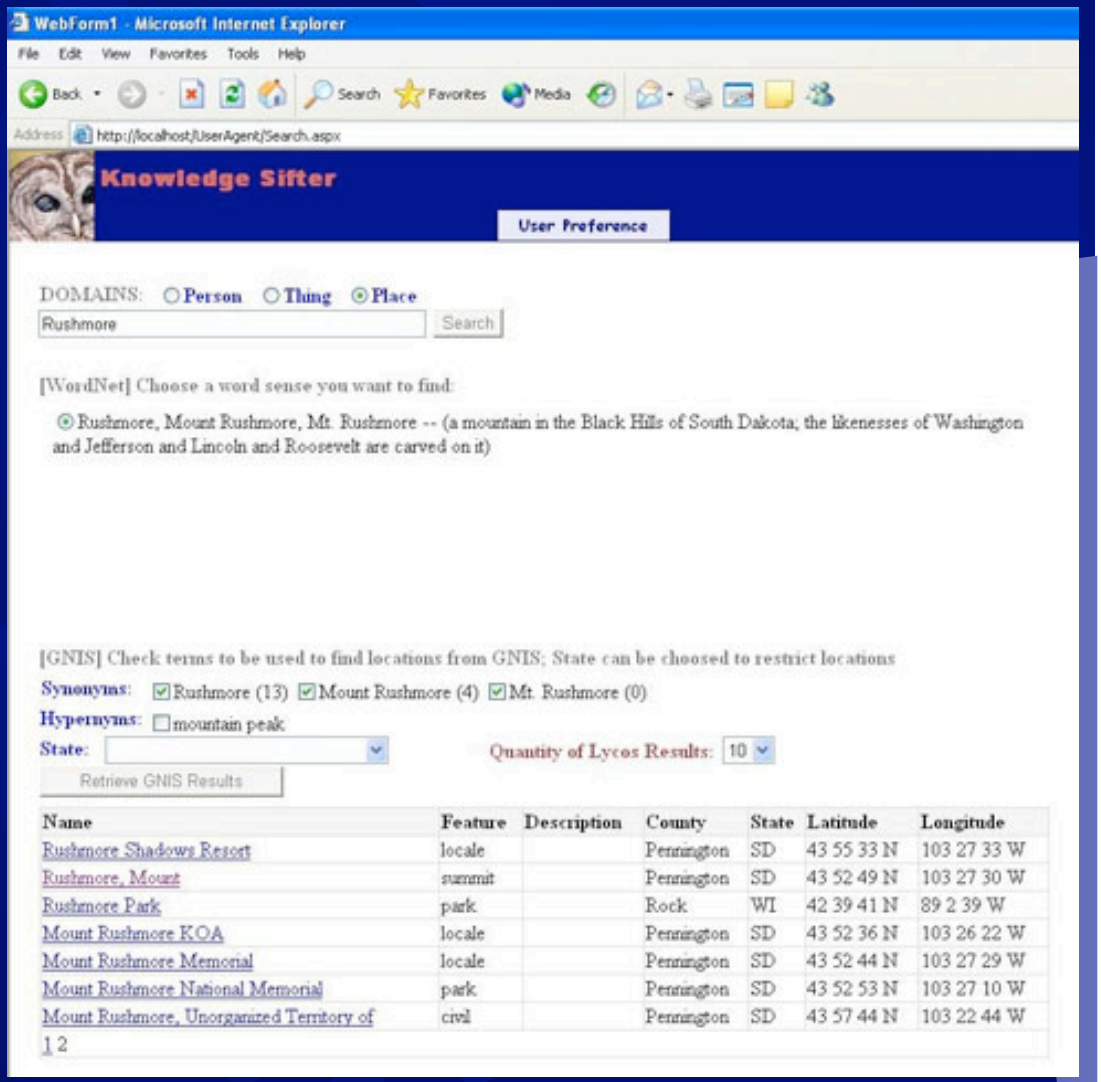
Ontology Taxonomy in OWL

- Ontology represents the conceptual model for images
- An Image has several Features such as Date and Size, with their respective attributes.
- An Image has Source and Content such as Person, Thing, or Place.
- Types are related by relationships and ISA relationships.
- Attributes of types are represented as properties.



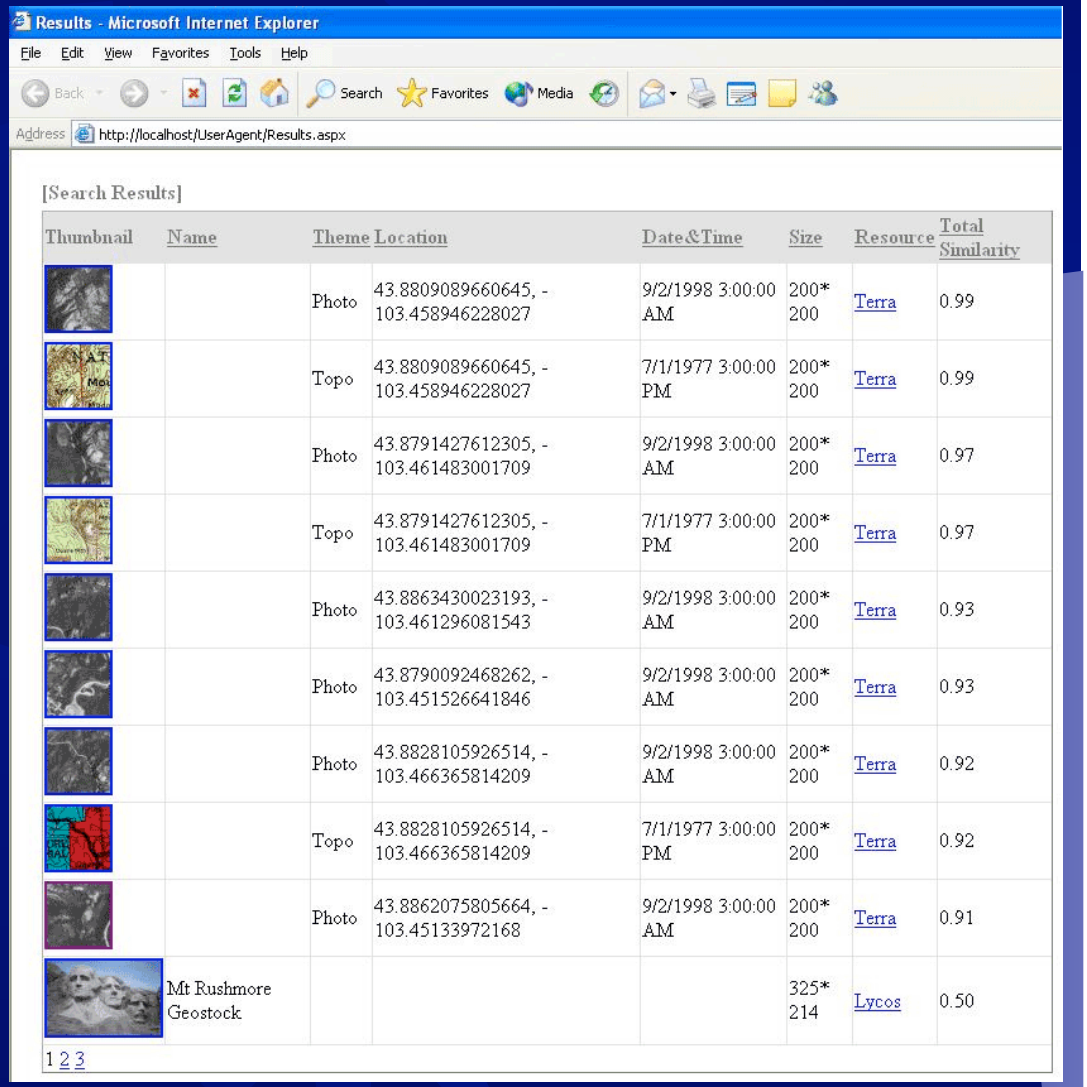
User Query Form

- User selects a Place and types 'Rushmore'
- WordNet provides related synonym concepts.
- GNIS is queried with synonyms to obtain latitude and longitudes for images
- Results from WordNet and GNIS are used to query the Lycos Images and TerraServer



KS Ranked Query Results

- Knowledge Sifter ranks search results according to user preferences
- Thumbnails allow user to browse the products and select appropriate images.



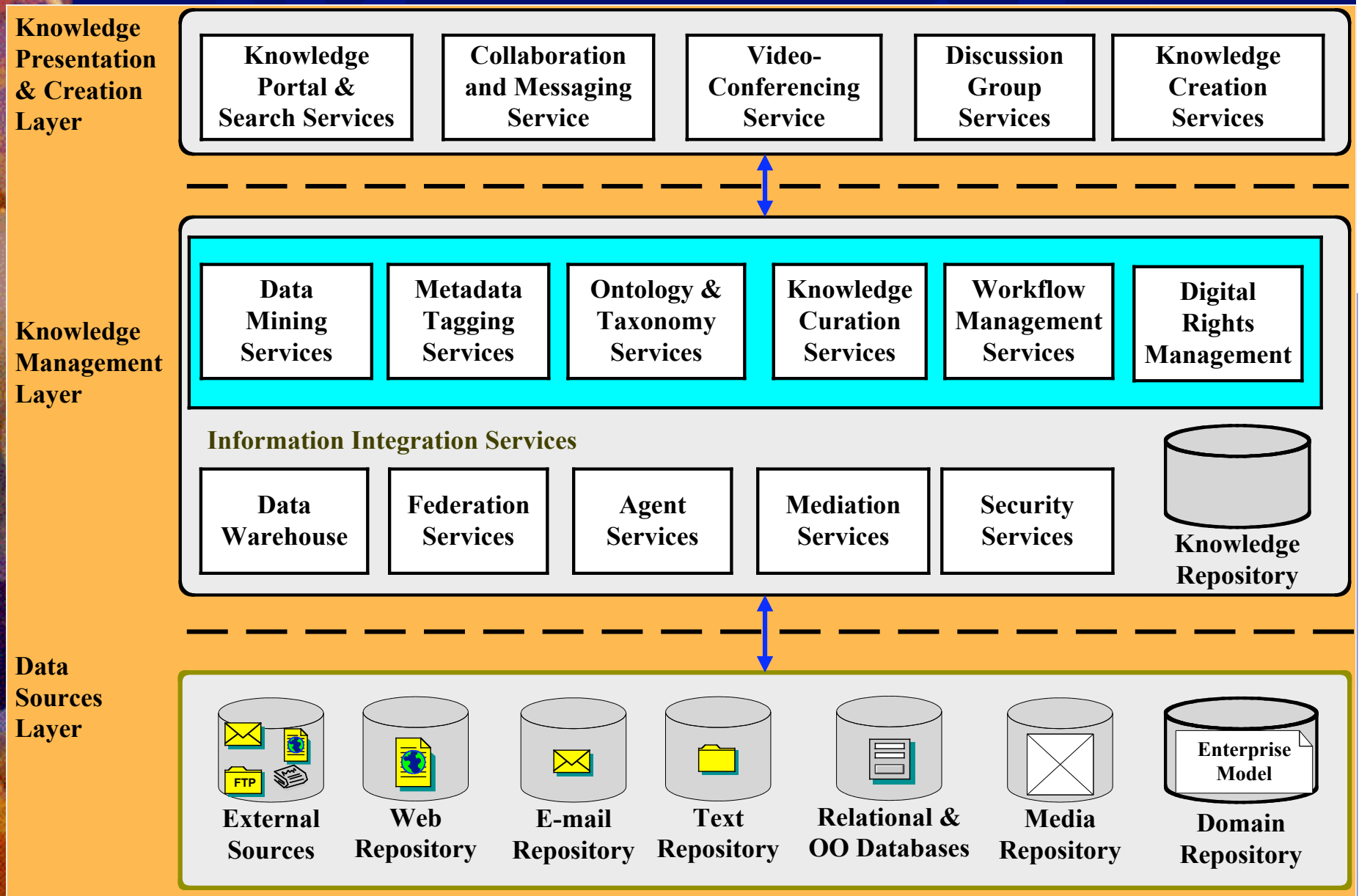
Thumbnail	Name	Theme	Location	Date&Time	Size	Resource	Total Similarity
		Photo	43.8809089660645, -103.458946228027	9/2/1998 3:00:00 AM	200*200	Terra	0.99
		Topo	43.8809089660645, -103.458946228027	7/1/1977 3:00:00 PM	200*200	Terra	0.99
		Photo	43.8791427612305, -103.461483001709	9/2/1998 3:00:00 AM	200*200	Terra	0.97
		Topo	43.8791427612305, -103.461483001709	7/1/1977 3:00:00 PM	200*200	Terra	0.97
		Photo	43.8863430023193, -103.461296081543	9/2/1998 3:00:00 AM	200*200	Terra	0.93
		Photo	43.8790092468262, -103.451526641846	9/2/1998 3:00:00 AM	200*200	Terra	0.93
		Photo	43.8828105926514, -103.466365814209	9/2/1998 3:00:00 AM	200*200	Terra	0.92
		Topo	43.8828105926514, -103.466365814209	7/1/1977 3:00:00 PM	200*200	Terra	0.92
		Photo	43.8862075805664, -103.45133972168	9/2/1998 3:00:00 AM	200*200	Terra	0.91
	Mt Rushmore Geostock				325*214	Lycos	0.50

1 2 3

Knowledge Sifter Conclusions

- **Knowledge Sifter has several interesting architectural properties:**
  - The architecture is service-oriented and provides intelligent middleware services to access heterogeneous data sources.
  - Line agents and staff agents cooperate to maintain services and knowledge bases
  - Ontology agent can consult multiple information sources to allow queries to be 'semantically enhanced'.
  - Agents are specified as Web services and use standard protocols such as SOAP, WSDL, UDDI, OWL.
  - New ontologies can be added by updating the OWL schema with new types and relationships
  - New data sources can be added by appropriately registering them with Knowledge Sifter.

Service-Oriented Knowledge Management Framework



Conclusions

- Organizational and technological trends suggest that *agent-based "intelligent middleware" services* can be used to provide knowledge management services over heterogeneous information sources
- Increasingly, organizations will create *dynamically configured virtual organizations* using Semantic Web services
- *Search and information integration services* are important components of a knowledge management strategy.

Publications

- Kerschberg, L. Functional Approach to in Internet-Based Applications: Enabling the Semantic Web, E-Business, Web Services and Agent-Based Knowledge Management. in Gray, P.M.D., Kerschberg, L., King, P. and Poulovassilis, A. eds. The Functional Approach to Data Management, Springer, Heidelberg, 2003, 369-392.
- Kerschberg, L., Knowledge Management in Heterogeneous Data Warehouse Environments. International Conference on Data Warehousing and Knowledge Discovery, (Munich, Germany, 2001), Springer, 1-10.
- Kerschberg, L., Chowdhury, M., Damiano, A., Jeong, H., Mitchell, S., Si, J. and Smith, S., Knowledge Sifter: Ontology-Driven Search over Heterogeneous Databases. (Submitted for Publication).
- Kerschberg, L., Gomaa, H., Menasce, D. and Yoon, J.P., Data and Information Architectures for Large-Scale Distributed Data Intensive Information Systems. Proceedings Eighth International Conference on Statistical and Scientific Database Management, (Stockholm, Sweden, 1996).
- Kerschberg, L., Kim, W. and Scime, A., Intelligent Web Search via Personalizable Meta-Search Agents. International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2002), (Irvine, CA, 2002).
- Kerschberg, L., Kim, W. and Scime, A. A Semantic Taxonomy-Based Personalizable Meta-Search Agent. in Truszkowski, W. ed. Workshop on Radical Agent Concepts (LNAI 2564), Springer-Verlag, Heidelberg, 2002.
- Kerschberg, L. and Weishar, D.J. Conceptual Models and Architectures for Advanced Information Systems. *Applied Intelligence*, 13. 149-164.
- Kim, W., Kerschberg, L. and Scime, A. Learning for Automatic Personalization in a Semantic Taxonomy-Based Meta-Search Agent. *Electronic Commerce Research and Applications* (ECRA), 1 (2).
- Menasce, D.A., Gomaa, H. and Kerschberg, L., A Performance-Oriented Design Methodology for Large-Scale Distributed Data Intensive Information Systems. First IEEE International Conference on Engineering of Complex Computer Systems, (Southern Florida, USA, 1995).
- Please visit the Publications section of the E-Center for E-Business Web site to download select publications.